How Google works

Christiane ROUSSEAU

Département de mathématiques et de statistique and CRM
Université de Montréal

June 9, 2010

1 Introduction

From its very beginning, Google became "the"search engine. This comes from the supremacy of its ranking algorithm: the *PageRank algorithm*. Indeed, with the enormous quantity of pages on the World-Wide-Web, many searches end up with thousands or millions of results. If these are not properly ordered, then the search may not be of any help, since no one can explore millions of entries.

Hoes does the PageRank algorithm work?

We will explain this. But, before let us make a search on Google. On June 4 2010, you get 16,300,000 results for *Klein project*, even though the project is just beginning. On that precise date, the first entry is

http://www.mathunion.org/icmi/other-activities/klein-project/introduction/

rather than

http://www.kleinproject.org/

The first url is the url of a *page*, which is located on the website of the International Mathematical Union: http://www.mathunion.org. Because the International Mathematical Union is an important body, its official website comes first when you make the search "International Mathematical Union". Moreover, it communicates some of its importance to all of its pages, one of which is

http://www.mathunion.org/icmi/other-activities/klein-project/introduction/

In a few months or years from now, we could expect that the page

http://www.kleinproject.org/

will appear first in a search for *Klein project*.

To explain the algorithm, we model the web as an oriented graph. The vertices are the *pages*, and the oriented edges are the *links* between pages. As we just explained, each page corresponds to a different url. Hence, a website may contain many pages. The model makes no difference between the individual pages of a website and its front page. But, most likely, the algorithm will rank better the front page of an important website.

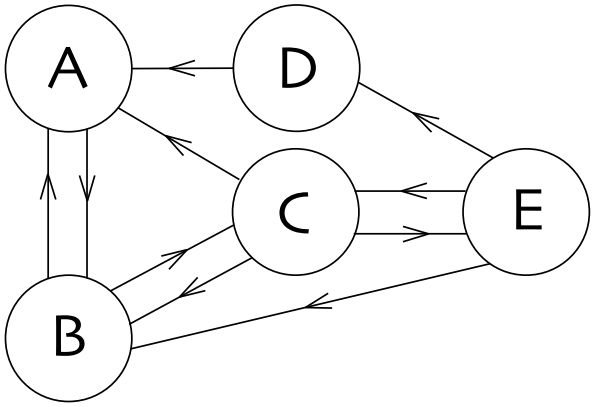

Figure 1: A simple web.

2 A simple example

Let us look at the simple web of Figure 1 with five pages named A, B, C, D, and E. This web has few links. If we are on page A, then there is only one link to page B, while, if we are on page C, we find three links, and we can choose to move to either page A, or B, or E. Remark that there is at least one link from each page.

We play a game, which is simply a *random walk* on the oriented graph. Starting from a page, at each step we choose at random a link from the page where we are, and we follow it. For instance, in our example, if we start on page B, then we can go to A or to C with probability $1/2$ for each case while, if we start on D, then we necessarily go to A with probability 1. We iterate the game.

Where will we be after n steps?

To automatize the process, we summarize the web in the following matrix P, where each column represents the departing page, and each row, the page where we arrive.

$$P = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

Let us remark that the sum of entries of any column of P is equal to 1 and that all entries are greater or equal to zero. Matrices having these two properties are very special: each such matrix is the matrix of a *Markov chain process*, also called *Markov transition matrix*. It always has 1 as an eigenvalue and there exists an eigenvector of 1, all the components of which are greater or equal to 1, with sum equal to 1. But, before recalling the definitions of eigenvalue and eigenvector, let us explore the advantage of the matrix representation of the web graph.

We consider a random variable X_n with values in the set of pages $\{A, B, C, D, E\}$, which contains N pages. X_n represents the page where we are after n steps of the random walk. Hence, if we call p_{ij} the entry of the matrix P located on the i-th row and j-th column, then p_{ij} is the conditional probability that we are on the i-th page at step $n+1$, given that we were on the j-th page at step n:

$$p_{ij} = \text{Prob}(X_{n+1} = i \mid X_n = j).$$

Note that this probability is independent of n! We say that a Markov chain process has *no memory of the past*.

It is not difficult to figure that the probabilities after two steps can be summarized in the matrix P^2. Indeed,

$$\begin{aligned}
\text{Prob}(X_{n+2} = i \mid X_n = j) &= \sum_{k=1}^{N} \text{Prob}(X_{n+2} = i \text{ and } X_{n+1} = k \mid X_n = j) \\
&= \sum_{k=1}^{N} \frac{\text{Prob}(X_{n+2} = i \text{ and } X_{n+1} = k \text{ and } X_n = j)}{\text{Prob}(X_n = j)} \\
&= \sum_{k=1}^{N} \frac{\text{Prob}(X_{n+2} = i \text{ and } X_{n+1} = k \text{ and } X_n = j)}{\text{Prob}(X_{n+1} = k \text{ and } X_n = j)} \frac{\text{Prob}(X_{n+1} = k \text{ and } X_n = j)}{\text{Prob}(X_n = j)} \\
&= \sum_{k=1}^{N} \frac{\text{Prob}(X_{n+2} = i \text{ and } X_{n+1} = k)}{\text{Prob}(X_{n+1} = k)} \frac{\text{Prob}(X_{n+1} = k \text{ and} X_n = j)}{\text{Prob}(X_n = j)} \\
&= \sum_{k=1}^{N} \text{Prob}(X_{n+2} = i \mid X_{n+1} = k) \text{Prob}(X_{n+1} = k \mid X_n = j) \\
&= \sum_{k=1}^{N} p_{ik} p_{kj} \\
&= (P^2)_{ij}.
\end{aligned}$$

In our example

$$P^2 = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{11}{18} \\ 0 & \frac{2}{3} & \frac{4}{9} & 1 & \frac{1}{9} \\ \frac{1}{2} & 0 & \frac{5}{18} & 0 & \frac{1}{6} \\ 0 & 0 & \frac{1}{9} & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{1}{9} \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

Iterating this idea, it is clear that the entry $(P^m)_{ij}$, of the matrix P^m describes the probability $\text{Prob}(X_{n+m} = i \mid X_n = j)$. For instance,

$$P^{32} = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0.293 & 0.293 & 0.293 & 0.293 & 0.293 \\ 0.390 & 0.390 & 0.390 & 0.390 & 0.390 \\ 0.220 & 0.220 & 0.220 & 0.220 & 0.220 \\ 0.024 & 0.024 & 0.024 & 0.024 & 0.024 \\ 0.073 & 0.073 & 0.073 & 0.073 & 0.073 \end{pmatrix} \end{array} \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}$$

All columns of P^{32} are identical, and the same as the columns of P^n when $n > 32$. Hence, after n steps, where n is sufficiently large, the probability of being on a page is independent from where we started!

Moreover, let us consider the vector

$$\pi^t = (0.293, 0.39, 0.22, 0.024, 0.0730)$$

(π is a vertical vector, and its transpose π^t is a horizontal vector). It is easily checked that $P\pi = \pi$. If we consider the i-th coordinate of the vector π^t as the probability of being on page i at a given time n, and hence π^t as the probability distribution of pages at time n, then it is also the probability distribution at time $n+1$. For this reason the vector π is called the *stationary distribution*. This stationary distribution allows to order the pages. In our example, we order the pages as B, A, C, E, D, and we declare B the most important page.

3 The general case

The general case can be treated exactly as our example. We represent the web as an oriented graph in which the N vertices represent the N pages of the web, and the oriented edges represent the links between pages. We summarize the graph in an $N \times N$ matrix, P, with the j-column representing the j-departing page and the i-row, the ith arrival page. In our example we found a vector π satisfying $P\pi = \pi$. This vector is an eigenvector of the eigenvalue 1. Let us recall

Definition 3.1 Let P be an $N \times N$ matrix. A number $\lambda \in \mathbb{C}$ is an *eigenvalue* of P if there exists a nonzero vector $X \in \mathbb{C}^N$ such that $PX = \lambda X$. Any such vector X is called an *eigenvector* of P.

We also recall the method to find the eigenvalues and eigenvectors.

Proposition 3.2 *Let P be an $N \times N$ matrix. The eigenvalues of P are the roots of the polynomial $\det(\lambda I - P) = 0$, where I is the $N \times N$ identity matrix. The eigenvectors of an eigenvalue λ are the nonzero solutions of the linear homogeneous system $(\lambda I - P)X = 0$.*

The following theorem of Frobenius guarantees that for the matrix associated to a web graph, we will always find a stationary solution.

Theorem 3.3 *(Frobenius) We consider an $N \times N$ Markov transition matrix $P = (p_{ij})$ (i.e. $p_{ij} \in [0, 1]$ for all i, j, and the sum of entries on each column is equal to i, namely $\sum_{i=1}^{N} p_{ij} = 1$). Then*

(i) $\lambda = 1$ is one eigenvalue of P.

(ii) Any eigenvalue λ of P satisfies $|\lambda| \leq 1$.

(iii) There exists an eigenvector X of the eigenvalue 1, all the coordinates of which are greater or equal than zero.

In our example, if we had taken any nonzero vector X with $X^t = (p_1, \ldots, p_N)$, where $p_i \in [0,1]$ and $\sum_{i=1}^N p_i = 1$, then we would have got that $\lim_{n\to\infty} P^n X = \pi$. Remark that the theorem does not guarantee that any matrix P satisfying the hypotheses of the theorem will have this property. Let us describe the possible pathologies and the remedy.

Possible pathologies.

- The eigenvalue 1 may be a multiple root of the characteristic polynomial $\det(\lambda I - A) = 0$.
- The matrix P may have other eigenvalues λ than 1, with modulus equal to 1.

What do we do in this case?

Remedy. We consider the $N \times N$ matrix $Q = (q_{ij})$, with $q_{ij} = \frac{1}{N}$ for all i, j. We replace the matrix P of the web by the matrix

$$P' = (1-\beta)P + \beta Q.$$

Remark that the matrix P' still has nonnegative entries and that the sum of the entries of each column is still equal to 1. We have the following result.

Theorem 3.4 *Given any Markov transition matrix P, there always exists a positive β, as small as we wish, such that all eigenvalues of the matrix P', except the eigenvalue 1, have modulus smaller than 1, and such that 1 is a simple eigenvalue (i.e. its multiplicity is 1). Let π be the eigenvector of 1, normalized so that the sum of its coordinates be 1. For such a matrix P', given any nonzero vector X, where $X^t = (p_1, \ldots, p_N)$ with $p_i \in [0,1]$ and $\sum_{i=1}^N p_i = 1$, then*

$$\lim_{n\to\infty} P^n X = \pi.$$

Another vignette will treat the Banach fixed point theorem. The theorem above can be seen as a particular application of it. Indeed,

Theorem 3.5 *We consider $\mathcal{S} = \{X \mid X^t = (p_1, \ldots p_N), p_i \in [0,1], \sum_{i=1}^N p_i = 1\}$, with the usual Euclidean distance. Then $\mathcal{S}$ is a complete metric space. On $\mathcal{S}$, we consider the linear operator $L : \mathcal{S} \to \mathcal{S}$ defined by $L(X) = P'X$, where P' is a matrix as in the previous theorem. The operator L is a contraction on $\mathcal{S}$, namely there exists $c \in [0,1[$ such that for all $X, Y \in \mathcal{S}$,*

$$\| L(X) - L(Y) \| \leq c \, \| X - Y \| .$$

Then, there exists a unique vector $\pi \in \mathcal{S}$ such that $L(\pi) = \pi$. Moreover, given any $X_0 \in \mathcal{S}$, we can define the sequence $\{X_n\}$ by induction, where $X_{n+1} = L(X_n)$. Then, $\lim_{n\to\infty} X_n = \pi$.

This theorem not only guarantees the existence of π, but gives a method to construct it, as the limit of the sequence $\{X_n\}$. We have seen an illustration of this convergence in our

example. Of course, in our example, we could also have found directly the vector π by solving the system $(I-P)X=0$ with matrix

$$I-P=\begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{3} & -1 & 0 \\ -1 & 1 & -\frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & -\frac{1}{2} & 1 & 0 & -\frac{1}{3} \\ 0 & 0 & 0 & 1 & -\frac{1}{3} \\ 0 & 0 & -\frac{1}{3} & 0 & 1 \end{pmatrix}$$

We wound have found that all solutions are of the form $(4s, \frac{16}{3}s, 3s, \frac{1}{3}s, s)^t$ for $s \in \mathbb{R}$. The solution whose sum of coordinates is 1 is hence π, where

$$\pi^t = \left(\frac{12}{41}, \frac{16}{41}, \frac{9}{41}, \frac{1}{41}, \frac{3}{41}\right).$$

References

[1] M. Eisermann, *Comment Google classe les pages webb*, http://images.math.cnrs.fr/Comment-Google-classe-les-pages.html, 2009.

[2] C. Rousseau and Y. Saint-Aubin, *Mathematics and technology*, SUMAT Series, Springer-Verlag, 2008 (A French version of the book exists, published in the same series.)